JSC-08453
NASA CR-
140214

EARTH RESOURCES PROGRAM

EVALUATION OF COMPUTER-AIDED PROCEDURE FOR DETECTING SURFACE WATER

SCIENCE AND APPLICATIONS DIRECTORATE

National Aeronautics and Space Administration

LYNDON B. JOHNSON SPACE CENTER

Houston, Texas

AUGUST 1973

JSC-08453

EVALUATION OF COMPUTER-AIDED PROCEDURE
FOR
DETECTING SURFACE WATER

Approved By:

R. B. Erb
Manager, Applications Office

R. B. MacDonald
Chief, Earth Observations Division

EARTH OBSERVATIONS DIVISION
SCIENCE AND APPLICATIONS DIRECTORATE
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
LYNDON B. JOHNSON SPACE CENTER
Houston, Texas

August 1973

PREFACE

The purpose of this document is to describe the evaluation of a maximum likelihood classification procedure for detecting and locating surface water using data from the multispectral scanner (MSS) on the Earth Resources Technology Satellite (ERTS-1), this activity was undertaken to support implementation of Federal legislation requiring the inventory of impoundments.

This document includes background information on why the valuation was conducted; a statement of the problem; a detailed description of the technical approach used; a statement of the performance results; and recommendations as to the most appropriate procedure to evaluate next.

This document was prepared pursuant to requirements identified with the Applications Office of the Earth Observation Division. It is comprised of a joint effort of personnel within the Earth Observations Division and personnel within the Earth Resources Department, Lockheed Electronics Company, Inc. Prime technical contribution to the effort described herein was made by T. C. Minter, scientific programming analyst. Activities by the contractor were authorized under contract NAS 9-12200.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ADP	Automatic Data Processing
CCT	Computer compatible tape
EOD	The Earth Observations Division of NASA at JSC
ERTS-1	Earth Resources Technology Satellite
ISOCLS	An Interative Self-Organizing Clustering (ISOCLS) Program which groups spectrally "similar" pixels into sets called clusters
DLMIN	Threshold value for combining clusters in ISOCLS
ISTOP	Maximum number of ISOCLS iterations
LARS-12 cards	Data cards used to define the training and test fields for ISOCLS
MAXCLS	The maximum number of clusters into which ISOCLS is allowed to cluster the data
NMIN	The minimum number of data points allowed per cluster in ISOCLS
STDMAX	The threshold for splitting clusters in ISOCLS
JSC	The Johnson Space Center in Houston, Texas
LARSAA	A computer software system used to classify and display multispectral scanner data on the basis of spectral information
CLASSIFY	A LARSAA processor which performs maximum likelihood classification on multispectral data.
DISPLAY	A LARSAA processor which allows a user to threshold pixels (i.e. assign a pixel to the unclassified category if it does not exceed a user specified confidence level) and display the classification results in the form of a line printer character map.
MAPTAP	An intermediate results tape produced by the LARSAA classification processor, CLASSIFY, which contains the identification of the class to which the pixel was assigned and its likelihood value.

ACRONYMS (cont.)

PICMON	A processor in LARSAA that generates a grey scale map on a line printer from a single channel of multispectral scanner data
MSS	Multispectral scanner on board the Earth Resources Technology Satellite
NASA	The National Aeronautics and Space Administration
NPID	The National Program for the Inspection of Dams
REGSTR	A computer program which correlates and registers multispectral images to a predetermined scale.

1.0 SUMMARY

In this document, a description is presented of the results from a detailed evaluation of a computer-aided procedure for processing ERTS-1 data to detect and locate surface water for the National Program for the Inspection of Dams (NPID). The procedure was evaluated using data from a study area in the vicinity of the Lake Somerville area in Washington County, Texas.

The procedure evaluated consisted of (1) selecting water training fields, (2) aggregating the training samples together and clustering them into unimodal clusters, (3) computing the mean vector and covariance matrix for each cluster, (4) classifying all of the study area into classes corresponding to the clusters using the maximum likelihood classifier, and (5) thresholding out the non-water pixels.[1]

Water training fields were selected from the ERTS-1 multispectral digital image without the aid of ground truth using a grey map of channel 4. This constraint and associated rationale are discussed in detail.

The result of the evaluation was that the use of the procedure failed to provide acceptable performance results. The success criteria established for this study was that 90% of all areas of surface water of 10 acres or more had to be correctly identified and located with a frequency of false detection of 10% or less. The result from the procedure evaluation was that 100% of all areas of surface water of 10 acres or more were correctly detected, but the frequency of false detection was approximately 96.8%.

[1]A pixel is the basic unit in image reconstruction from the digital tape, using electronic display devices. It is the binary integer recorded on magnetic tape that represents a time sample of the analog scan line trace, the value of which is proportional to the energy sensed by each ERTS-1 MSS channel.

It was concluded from this evaluation that two principal sources of error existed; first, the water training field selection procedure potentially included pixels from the perimeter of areas of surface water, wet fields, and bare soil areas. The presence of the wet fields/ bare soil/perimeter pixels among the water training samples resulted in the presence of two clusters in the non-water part of feature space. These two clusters resulted in many false detections when wet areas and bare soil were classified as water. The development of a method for selecting thresholds which would eliminate most of the pixels assigned to these two clusters was beyond the scope of this evaluation. Secondly, another possible source of error was contributed because the ISOCLS clustering routine, as used in this evaluation, did not define unimodal clusters and compute meaningful statistics (mainly the covariance matrix) for each class. The classes defined by ISOCLS did not appear to conform to the normality assumptions and were generally multimodal. No conclusion was reached as to the effect of this problem on classification accuracy.

It is recommended as a follow-on to this exercise that (1) using representative and verified training samples for water only (i.e. water training samples that have been verified from photography), a procedure be defined for identifying water using LARSAA (with and without the aid of clustering), (2) using representative and verified water and non-water training samples a procedure be defined for identifying water using LARSAA (with and without the aid of clustering), (3) using representative and verified water and non-water training samples, train another classifier (such as described in Reference 3) which is independent of certain of the assumptions made in Gaussian maximum likelihood classification (such as the normality assumption and the unimodal classes assumption) for the purpose of obtaining an independent assessment of how well the maximum likelihood classifier is performing in meeting the assumptions on which the classifier is based, and (4) define a procedure for selecting representative training samples for water and non-water which does not include pixels from the perimeter of areas of surface water, wet fields and bare soil areas in the water training sample and does not include water pixels in the non-water training sample.

2.0 INTRODUCTION AND BACKGROUND

In this technical report the results from an evaluation of a procedure for processing of ERTS-1 data to detect and locate surface water in support of the National Program for the Inspection of Dams (NPID) will be discussed. The procedure was evaluated on data from a study area near Lake Somerville in Washington County, Texas. In this document the evaluated data processing procedure will be outlined and the results obtained will be discussed. An analysis of the results obtained at each step of the procedure will then be given. Conclusions drawn from this analysis will be given and recommendations as to the most appropriate procedure to evaluate next will be given.

For the purposes of understanding why this procedure was investigated, background related to the NPID is presented below.

In August 1972, the President signed into law Public Law 92-367 which authorized the Secretary of the Army to undertake a national program for the inspection of dams. The need for dam safety w[illegible] [illegible]ght to national attention when water impoundments in West Virg[illegible] and South Dakota gave way, resulting in loss of life and property.

In brief the law directs the Secretary of the Army, acting through the Chief of Engineers, to carry out a national program for the inspection of dams. The scope of water impoundment capacity covered by the law is graphically described in Figure 1. To determine whether a dam (including the waters impounded by such dam) constitutes a danger to human life or property, the Secretary will take into consideration the possibility that the dam might be endangered by overtopping, seepage, settlement, erosion, sediment, cracking, earth movement, earthquakes, failure of bulkheads, flashboard, gates on conduits, or other conditions which exist or which might occur in any area in the vicinity of the dam.

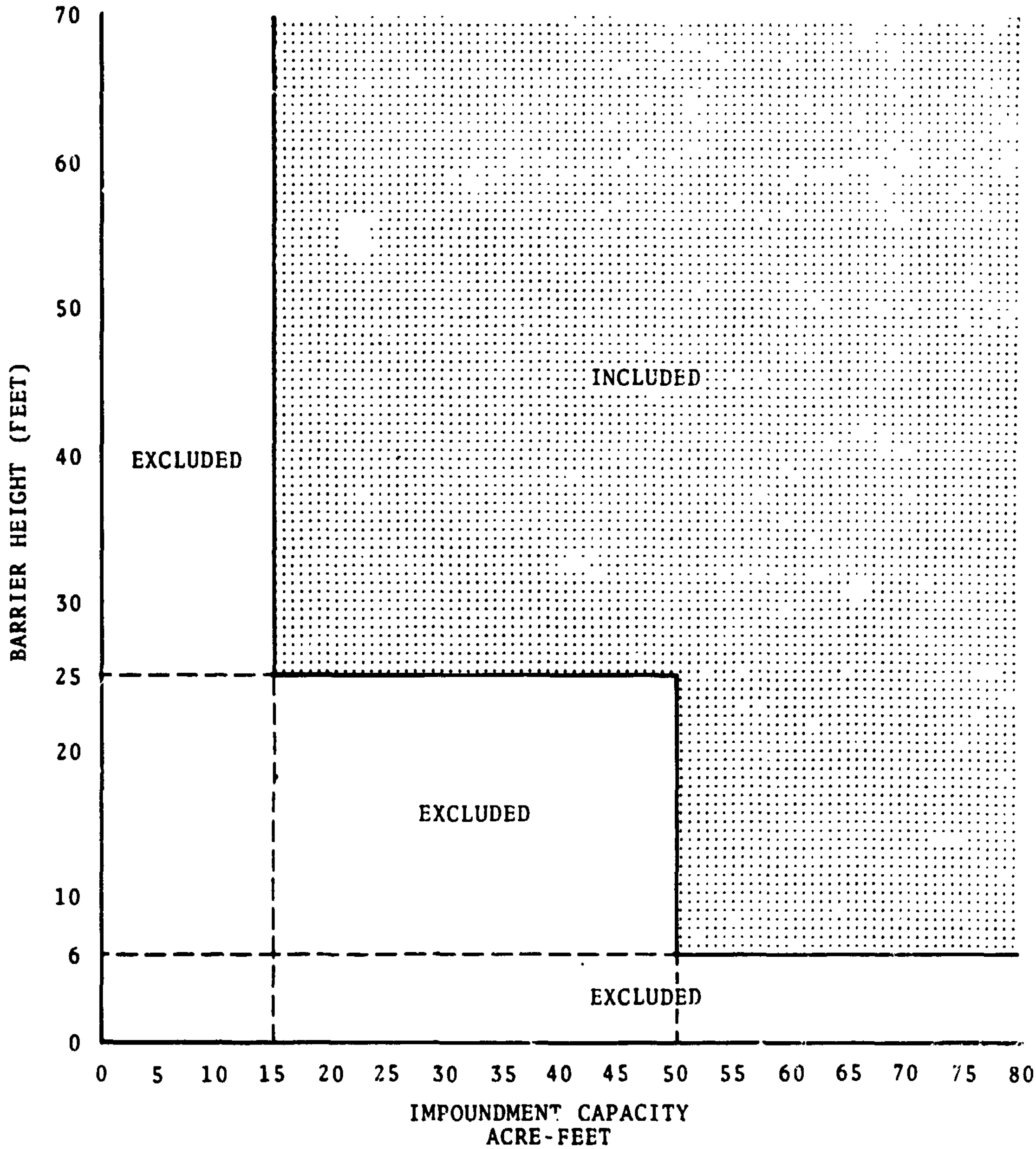


Figure 1. - Scope of water impoundment capacity included in Public Law 92-367.

The report by the Secretary of the Army to Congress is due on or before July 1, 1974. The report will include (1) an inventory of all dams of interest located in the United States, (2) a review of each inspection made, the recommendations furnished to the governor of the state in which such dam is located and information as to the implementation of such recommendations, and (3) recommendations for a comprehensive national program for the inspection, and regulation for safety purposes of dams of the nation, and the respective responsibilities which should be assumed by the federal, state, and local governments and by public and private interests.

In December, 1972, the Texas Water Rights Commission (TWRC) submitted, through the office of the Governor of Texas, a request for assistance by NASA/JSC/EOD in the development of a procedure or procedures for utilizing data acquired by the Earth Resources Technology Satellite (ERTS-1) in detecting and locating water impoundments.

The ERTS-1 satellite was placed in orbit to gather data relative to the environment of the Earth. The ERTS-1 satellite orbits the Earth in a circular, sun-synchronous, near-polar orbit at an altitude of approximately 494 nautical miles. The satellite orbits the earth approximately 14 times each day and views the same scene on the earth approximately every 18 days.

On board ERTS-1 is a multispectral scanner (MSS) which receives spectral information in four channels covering the following wavelengths:

Channel	Spectral Band	Wavelength (micrometers)	
1	4	0.5 - 0.6	visible
2	5	0.6 - 0.7	
3	6	0.7 - 0.8	reflective infrared
4	7	0.8 - 1.1	

The MSS data are recorded and transmitted in digital format to the NASA data processing facility. Two image products are produced for each scene, in the form of photographic images and digital images.

The MSS products used for the procedure described in this document are: (1) 9 1/2" x 9 1/2" system corrected photographic images at a scale of 1:1,000,000 and, (2) system corrected computer compatible tape (CCT) digital images. Each digital image consists of four CCT's, each CCT covering a strip 25 by 96.3 nautical miles.

3.0 GENERAL DESCRIPTION OF THE PROCEDURE EVALUATED

3.1 Statement of the Problem

Previous attempts to use the LARSAA maximum likelihood classifier program to detect surface water using the ERTS-1 data (Reference 2) met with limited success. The three procedures tested consisted of applying the maximum likelihood classifier by (a) selecting <u>water only</u> training fields, clustering the training samples, classifying, and then thresholding, (b) selecting <u>water</u> training fields and a <u>similar</u> class training fields, clustering both the water and similar class training samples, classifying, and then thresholding, and (c) selecting <u>water only</u> from known water bodies larger than 20 surface acres clustering the training samples, classifying, and then thresholding.

The results obtained from a test of procedure (a) (CLASSIFY maximum likelihood classification using water only training fields) were that the correct identification of 81% of areas of surface water of 10 acres or more, and a frequency of false detection of 82.5% were achieved.

Procedure (b) (CLASSIFY maximum likelihood classification using water plus a similar class training fields) resulted in the correct identification of 94% of all areas of surface water of 10 acres or more and had a frequency of false detection of 66%.

Procedure (c) (CLASSIFY maximum likelihood classification using only water bodies greater than 20 surface acres as training fields) correctly identified 69% of all areas of surface water of 10 acres or more, and had a frequency of false detection of 88%. None of these results were acceptable according to the established performance criteria. These established success criteria required a correct identification of 90% of all areas of surface water of 10 acres or more, and a frequency of false detection of 10% or less.

3.2 Procedure Evaluated

As a result of the evaluation of these procedures it was concluded that the best approach for identifying water would be to (1) select water training fields, (2) lump the training samples together and cluster them into unimodal clusters, (3) compute the mean vector and covariance matrix for each cluster, (4) classify all of the study area into these clusters using the maximum likelihood classifier and then, (5) threshold out the non-water pixels.

The procedure to be evaluated is similar to one of the procedures tested previously, but with two important modifications. First, the clustering routine ISOCLS, was to be set up to cluster the water training samples into a larger number of smaller clusters. The smaller clusters would help insure that the clusters were unimodal and conform closer to the normality assumption. Second, a systematic approach to selecting class thresholds, as described in Reference 1 was to be used. The clustering and the threshold selection procedure will be discussed in greater detail later in this document.

In testing this procedure an important constraint was imposed that required that all water training fields had to be identified on the ERTS-1 imagery without the aid of ground truth. Previous experience indicated that water had two important characteristics that might facilitate identification for training purposes. First, adjacent water pixels tend to have similar radiance values. Second, most water pixels have radiance values from 0 to 6 in channel 4 and any surrounding wet areas generally have radiance values of 7 to 9 in channel 4.

The water training field selection procedure then had the requirement to find these homogeneous areas of low radiance values of at least 8 samples in channel 4 on a LARSAA/PICMON grey map and use them as water training fields. To obtain a representative training sample of the various areas of surface water and to avoid biasing the water class statistics in favor of the large water bodies (whose location are generally known already), the number of training samples taken from any one area of surface water was restricted to a maximum of 30 samples.

The training samples were next aggregated together and clustered to obtain the water subclasses inherent in the data. by a proper setting of the clustering program's control parameters, the water training samples were partitioned into a large number of small clusters to insure the clusters were unimodal, and thus conform more closely to the established normality assumption which is the basis of the Gaussian maximum likelihood classifier.

Next, all of the study area data was classified, using the maximum likelihood classifier into the subclasses defined by ISOCLS. The maximum likelihood classification of water into the various water subclasses was not helpful in separating water from non-water. Thresholding was used to eliminate the non-water pixels from each water subclass. The thresholds effectively defined the discriminant boundary between water and non-water. The assumption made in using this procedure was that if the water subclasses were sufficiently well separated from the non-water classes in spectral space, then the thresholds would be an effective means of defining the discriminant boundary between water and non-water. It should be noted, however, that this is <u>not</u> a maximum likelihood classification rule. Maximum likelihood classification implies that a conditional probability density function is available for the non-water class and each pixel is assigned to the class (i.e. water or non-water) for which the conditional class probability is greatest. More precisely, the procedure described above for identifying water is a Gaussian hypothesis testing procedure where a pixel is assumed to be water if its conditional probability is greater than a certain confidence level. However, since the CLASSIFY maximum likelihood classifier was used to evaluate this procedure, the procedure will be referred to as a maximum likelihood procedure in this document.

A systematic approach to selecting class thresholds was used, and is described in detail in Appendix B. This is essentially the same as the procedure reported in Reference 1. The procedure will be described briefly.

In using the LARSAA processor routine DISPLAY, the investigator can specify the threshold value, T_i, for each material. In the past this was done in two ways.

(2) In the first proceudre the program DISPLAY was run with various values of T_i until the classification map (in the form of a line printer listing) appeared acceptable. This procedure is very subjective and is difficult to repeat.

(b) The second procedure is based on the accepted fact that if X (a data value) is distributed according to the multivariate normal distribution

$$P_i(X) = \frac{1}{(2\pi)^{N/2}|K_i|^{1/2}} e^{-1/2\, Q_i(X)} \qquad (1)$$

where

$$Q_i(X) = (X - M_i)^T K_i^{-1}(X - M_i) = 2T_i \qquad (2)$$

then the quantity $Q_i(X)$ in equation (2) is a random variable having a chi-square distribution with N degrees of freedom (where N is the dimension of the measurement vector X). In equation (2) M_i is the mean vector, K_i is the covariance matrix, and T_i is the threshold value for class i. To the extent that the training data is normally distributed the investigator can look up in a cumulative chi-square table the threshold value which will reject (i.e., assign to the unclassified category) no more than a specified percentage of samples. For example, a threshold setting of $T_i = 3.0$ will reject no more than 5 percent of samples drawn from a two-dimensional multivariate normal distribution.

In many cases the training data is not normally distributed and the distribution of Q_i must be determined empirically to select the threshold values. To empirically form an estimate of the <u>density</u> function of Q_i, the number of occurrences of each value of Q_i, among the training samples for class i, are counted. Then the density functions are accumulated (i.e. integrated) with respect to Q_i to form the <u>distribution</u> of Q_i for each class. Class thresholds T_i are then selected from the empirical distribution of Q_i where $T_i = 1/2Q_i$.

The specific procedure evaluated will be briefly described next. A detailed description of this procedure is given in Appendix A.

(1) The appropriate ERTS-1 data tape was selected and registered to a 1:24,000 scale map of the test area using the program REGSTR.

(2) The PICMON program was used to obtain a grey map of channel 4. On the grey map counts 0 to 6 were represented by the symbol M, counts 7 to 9 an asterisk (*), and counts 10 to 63 by a blank (no symbol).

(3) From the PICMON map, all areas were located that contained eight or more contiguously printed M or * symbols, and where at least 25% of the symbols were M symbols. Within these areas, the largest possible rectangle containing at least 50% M symbols was inscribed. If the resulting rectangle contained less than eight symbols (including M and *), the area was deleted from further consideration. If the rectangle consisted of more than 30 symbols (M's and *'s), the size of the rectangle was reduced in size, to contain no more than 30 symbols.

(4) All of the training samples obtained in the step above were aggregated together and clustered with ISOCLS, using channels 1 and 4 with MAXCLS = 20, NMIN = 16, STDMAX = 1.5, DLMIN = 3.2, and ISTOP = 20. Statistics were then obtained for each cluster.

(5) The study area near Lake Somerville was classified using channels 1 and 4 and using the cluster statistics obtained in step 4.

(6) The density and cumulative distribution of the quadratic form were estimated for each subclass. Threshold values were obtained from these cumulative distributions of the quadratic form (see Appendix B).

(7) DISPLAY was run with the threshold values obtained above in step 6.

(8) On the resultant display map, all areas representing class 0 (non-water areas of 10 surface acres or more improperly classified as water), Class III areas (10 or more surface acres), were located and evaluated against existing ground truth data. The established performance criteria for this evaluation is discussed in detail in Appendix C.

In the following sections, the results of the evaluation of the procedure just discussed will be documented. The analytical approach used to obtain these results will then be discussed. Conclusions arrived at will be presented, and recommendations for further actions will be presented.

4.0 RESULTS, CONCLUSIONS, AND RECOMMENDATIONS

4.1 Performance Test Results

The procedure described in section 3.2 was used to classify ERTS-1 data (ERTS-1 Scene: E-1092-16305) acquired on October 23, 1973. Approximately 500 areas were identified as Class III areas (refer to Appendix C). Of the 16 Class III areas determined from photographic ground truth (Mission 220, flown 8 November 1972), all such areas corresponded to identified Class III areas on the data display map. However, approximately 484 of the identifications of Class III areas corresponded to areas where no surface water was found to exist from photointerpretative "ground truth". Thus by the definitions of the performance criteria in Appendix C, the percentage of Class III areas correctly located by this procedure was

$$F_{33} = \frac{16}{16} \times 100 = 100\%$$

The frequency of false detection F_{03} was, however,

$$F_{03} = \frac{484}{16+484} = 96.8\%$$

Therefore, the chosen procedure failed to meet the criteria of a frequency of false detection of 10% or less, though it exceeded criteria of the correct identification of Class III areas of 90% or greater.

4.2 Analysis of the Results

Seven procedural steps were taken to develop the analytical results: The results of each step of the procedure described in the previous section are presented here along with an analysis of each step.

(A) Step 1 - An ERTS-1 data tape (ERTS-1 Scene: E-1092-16305) of the Lake Somerville Study Area was registered to a scale of 1:24,000 using REGSTR. A pixel dropout rate of approximately 25% was experienced in registering these data. This dropout rate may result in Class III areas being classified as Class I or II areas.

(B) Step 2 & 3 - These two steps involved the random selection of 20 water training fields. Only 20 fields were selected because this number is the maximum number of distinct rectangular areas that ISOCLS can process at one time.

(C) Step 4 - ISOCLS was used to cluster the 248 pixels from the 20 water training fields into 9 clusters, and statistics were defined using ISOCLS. The scatter plot in Figure 2 was computed for channels 1 and 4. The mean for each cluster is shown on the scatter plot in Figure 2. Table 1 lists the means vector and covariance matrix computed for each cluster.

On the basis of the scatter plots, no conclusions could be made as to the adequacy of ISOCLS for defining clusters and cluster statistics that could be used in a Gaussian maximum likelihood classification. In Figure 2, it can be noted that the cluster means generally fall near the modes of the data. A possible problem is noted with the covariances computed for some of the clusters: Clusters 1, 3, 6, 7, 8, and 9 had negative covariances. Negative covariances are normally not present in distributions encountered in remotely sensed data. The presence of negative covariances accentuated the problem of selecting thresholds. The negative covariances effectively moved the wet areas and vegetation closer to the means of the water clusters and made thresholding more difficult.

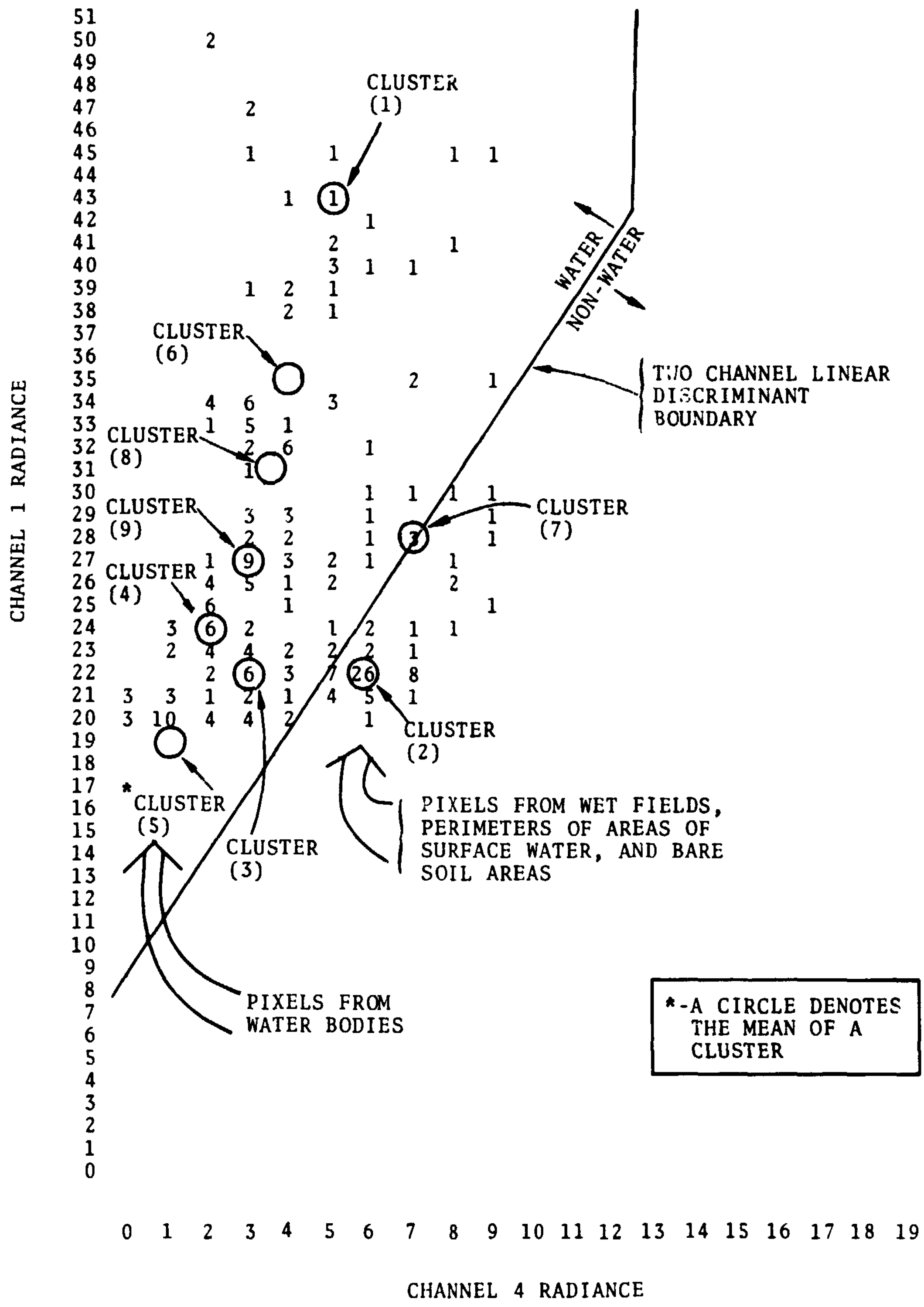


Figure 2. - Scatter plot of water training samples.

TABLE 1 — CLUSTER STATISTICS

MEANS

CLUSTER	CH(1)	CH(4)
1	43.20	5.05
2	22.05	5.98
3	21.63	3.19
4	24.32	1.96
5	18.64	.84
6	36.50	5.17
7	28.07	7.64
8	32.96	3.19
9	27.24	3.48

COVARIANCE MATRIX FOR CLUSTER 1

10.96	
-3.21	3.55

COVARIANCE MATRIX FOR CLUSTER 2

.63	
.10	.47

COVARIANCE MATRIX FOR CLUSTER 3

1.12	
-.01	.37

COVARIANCE MATRIX FOR CLUSTER 4

.93	
.15	.39

COVARIANCE MATRIX FOR CLUSTER 5

30.39	
1.26	.45

COVARIANCE MATRIX FOR CLUSTER 6

4.25	
-2.08	2.64

COVARIANCE MATRIX FOR CLUSTER 7

2.64	
-.40	1.23

COVARIANCE MATRIX FOR CLUSTER 8

.85	
-.47	.74

COVARIANCE MATRIX FOR CLUSTER 9

1.03	
-.03	.55

TOTAL NUMBER OF CLUSTERS = 9
TOTAL NUMBER OF POINTS = 248

CLUSTER	SYMBOL	POINTS IN CLUSTER
1	1	20
2	2	62
3	3	27
4	4	28
5	5	25
6	6	12
7	7	14
8	8	27
9	9	33

The most critical problem depicted by Figure 2 is the presence of a large number of pixels from the perimeter of areas of surface water, wet areas, and bare soil which are incorrectly included in water training fields by using the water training samples selection procedure described in Step 3 of the procedure.

On reviewing the water training fields within the Lake Somerville study area, it was found that approximately 40% of the water training fields selected corresponded to wet fields or bare soil areas as determined from photointerpretation "ground tr   The criteria used in Steps 2 and 3 also include what is belie   be some pixels related to the perimeter of an area of surface water. These pixels are partitioned into clusters 2 and 7. This was also evident from the scatter plot in Figure 2. Many of these wet area/bare soil/perimeter pixels fall on the non-water side of the Two Channel Linear Discriminant Boundary shown in Figure 2. This is an empirically derived discriminant boundary (see Reference 2) which has been shown to provide acceptable classification results (as described in Section 3.0, Appendix C), using ERTS-1 data for East Texas. Additionally, in subsequent classification results (see Step 5), these perimeter pixels, wet areas, and bare soil areas were identified as belonging to clusters 2 and 7.

(D) Step 5 - The Lake Somerville study area was classified using channels 1 and 4 and the cluster statistic obtained in Step 3. A LARSAA map tape (MAPTAP) was generated.

(E) Step 6 - The threshold values for each of the 9 subclasses were selected. An empirical threshold selection procedure (described in Appendix B), was used to select threshold values. This procedure provides a method for obtaining thresholds even when the classification data does not conform to the normality assumption.

The procedure described in Appendix B was used to form an empirical estimate of the density function of Q_i (see equation (2) page 2-4 by counting the number of occurrences of each value of Q_i among the training samples for class i. The density function of Q_i for each of the nine water subclasses are shown in Figure 3.

The density functions of Q_i were then summed with respect to Q_i to form the distribution of Q_i for each of the nine water subclasses as shown in Figure 4.

In addition, an empirical estimate of the density function of Q_i was formed by counting the number of occurrences of each value of Q_i among all of the pixels assigned to each of the nine water subclasses in the Lake Somerville study area. These density function estimates for the nine water subclasses are shown in Figure 5.

Two conclusions can be drawn from data presented in Figures 3 and 5; First, in Figure 3, the density and distribution function estimates of $Q_i(X)$ of the training samples for each of the nine water classes differed significantly from the theoretical chi-square density function and the theoretical cumulative chi-square distribution function shown in Figure 6. It can be concluded from the density function estimate of Q_i (Figure 3) for the training samples that in general the water class training samples are multimodal and do not conform to the normality assumption. The multimodality of the water training samples could possibly be attributed to the small number of training samples used to form the estimate of Q_i, but in Figure 5 the density functions of Q_i of all of the pixels assigned to each class in the Lake Somerville Study Area also differ from the theoretical chi-square density for low values of Q_i (i.e. for values of Q_i less than approximately 3 to 5). Since the non-water classes (i.e. vegetation, wet area, etc.) appear out in the tails of the density function of Q_i (i.e. for Q_i values greater than

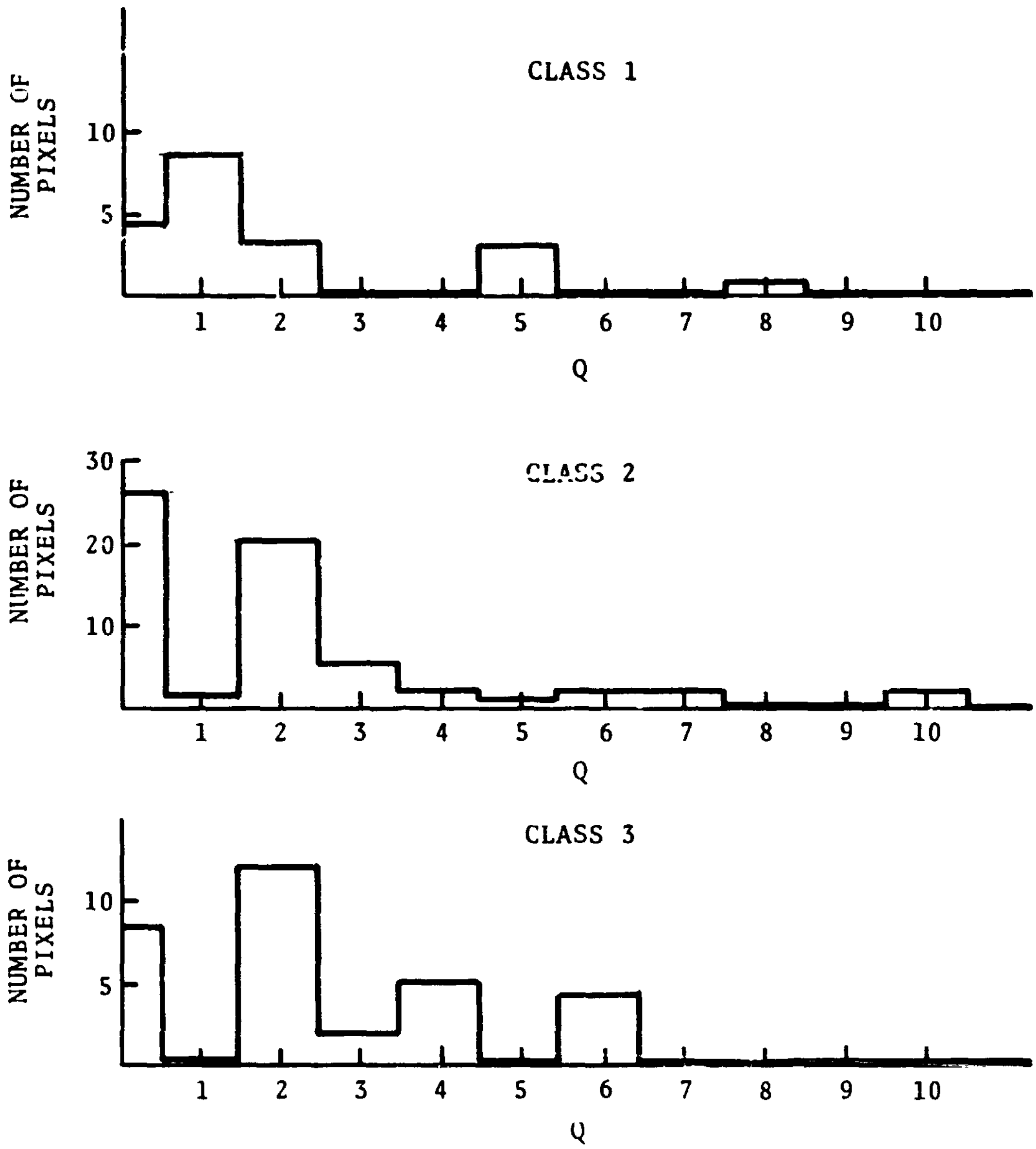


Figure 3. - Empirical estimate of the density function of the quadratic form Q_i for the nine water classes - estimated from the training samples.

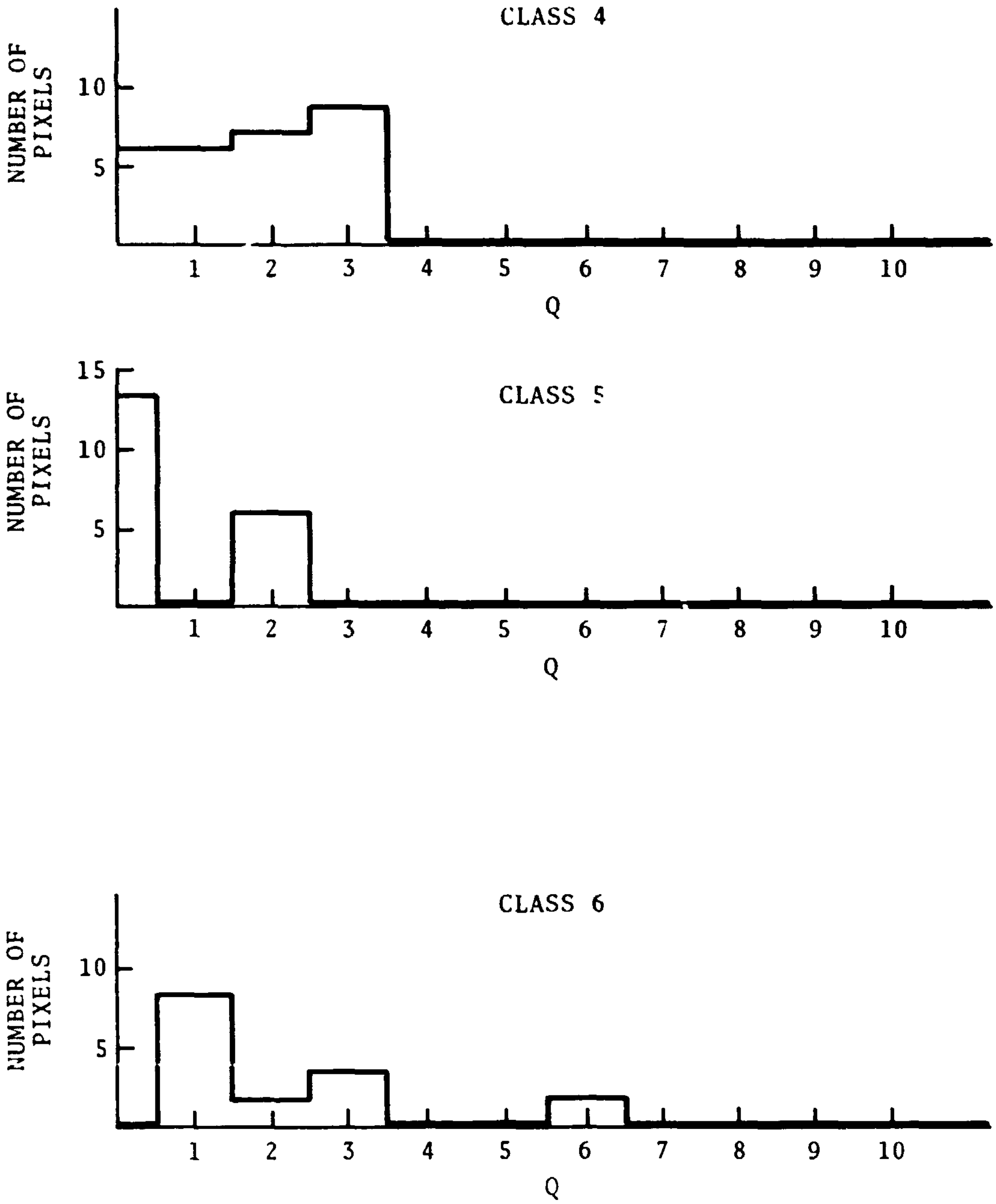


Figure 3. - Continued.

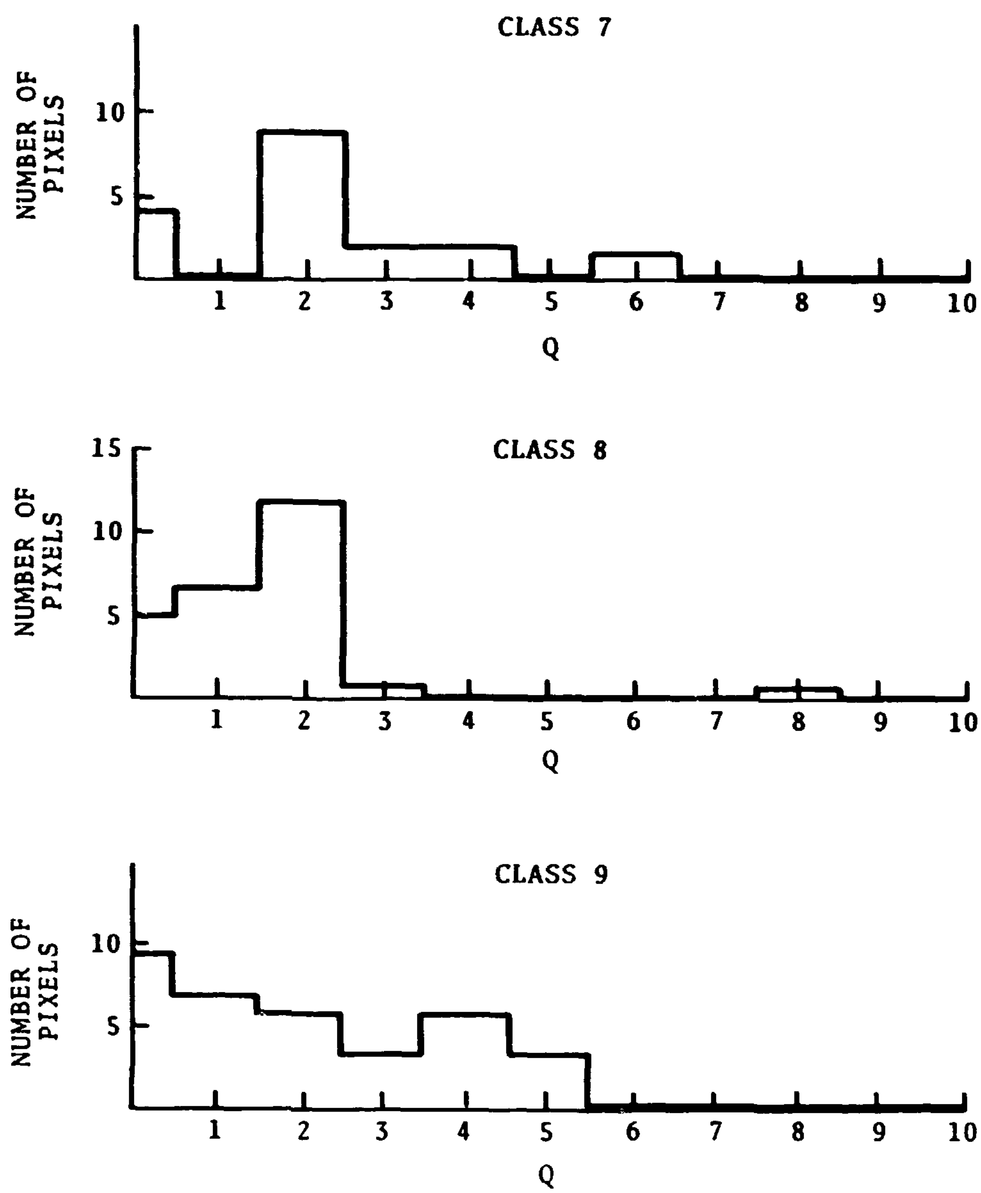


Figure 3. - Concluded.

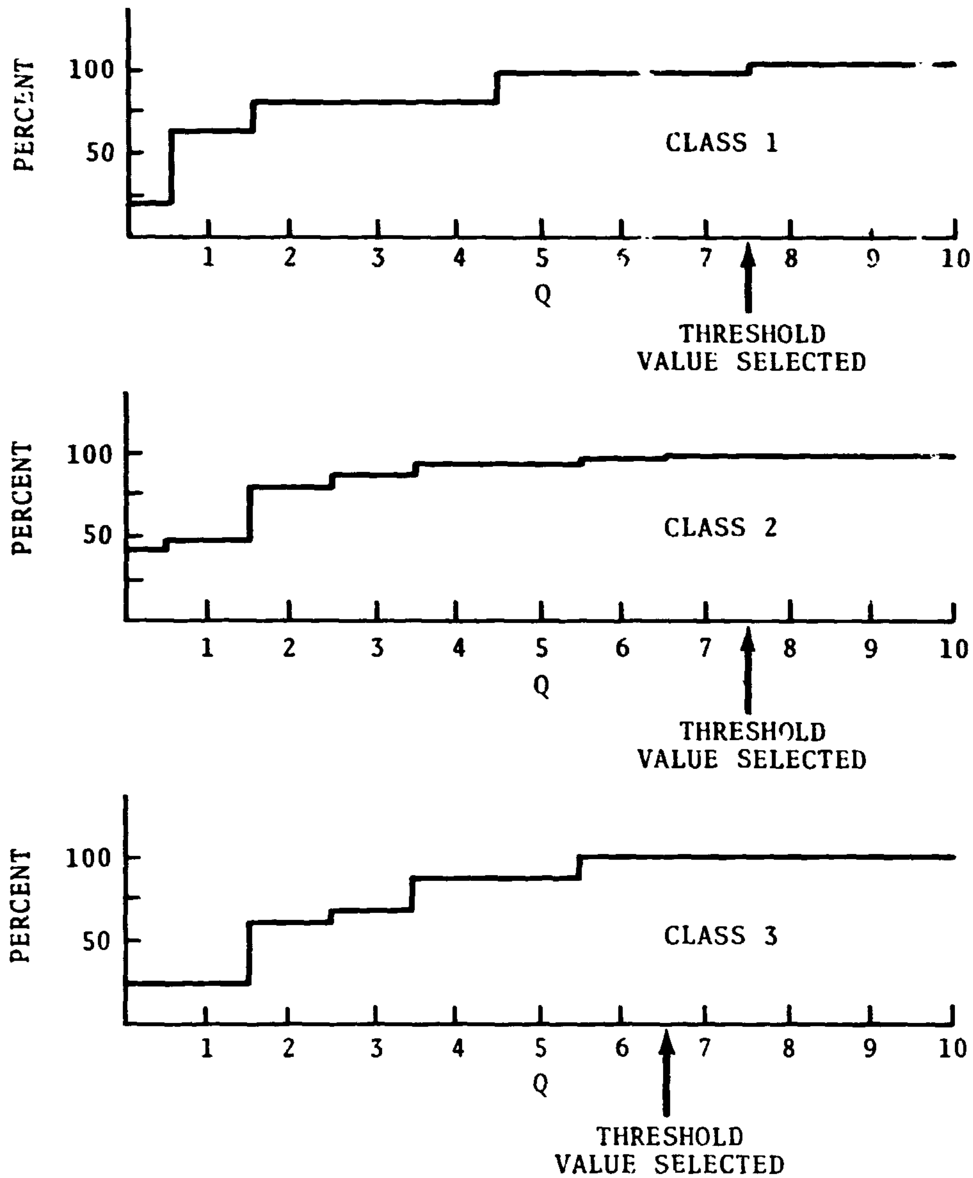


Figure 4. - Empirical estimate of the cumulative distribution of the quadratic form Q_i for the nine water classes - estimated from the training samples.

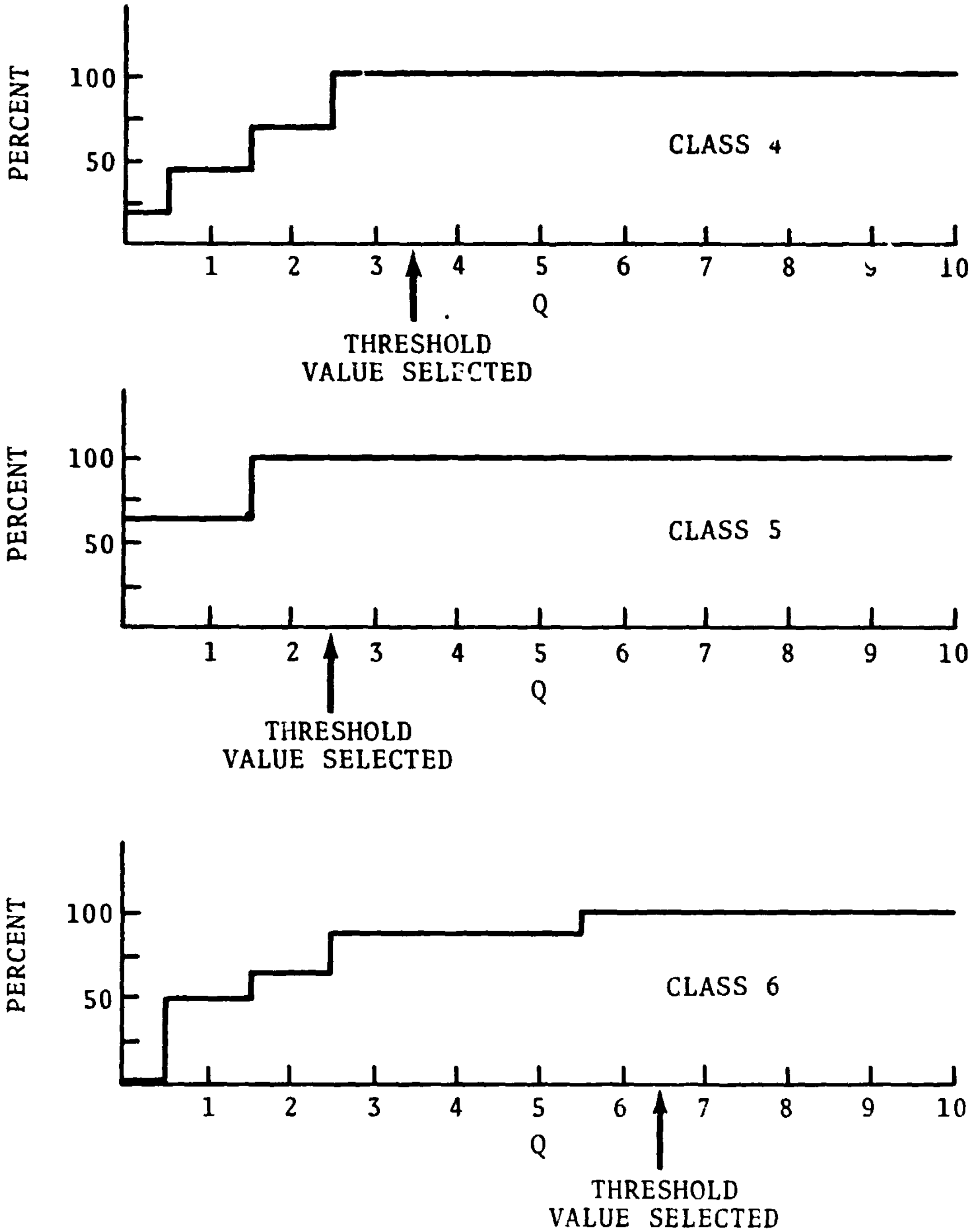


Figure 4. - Continued.

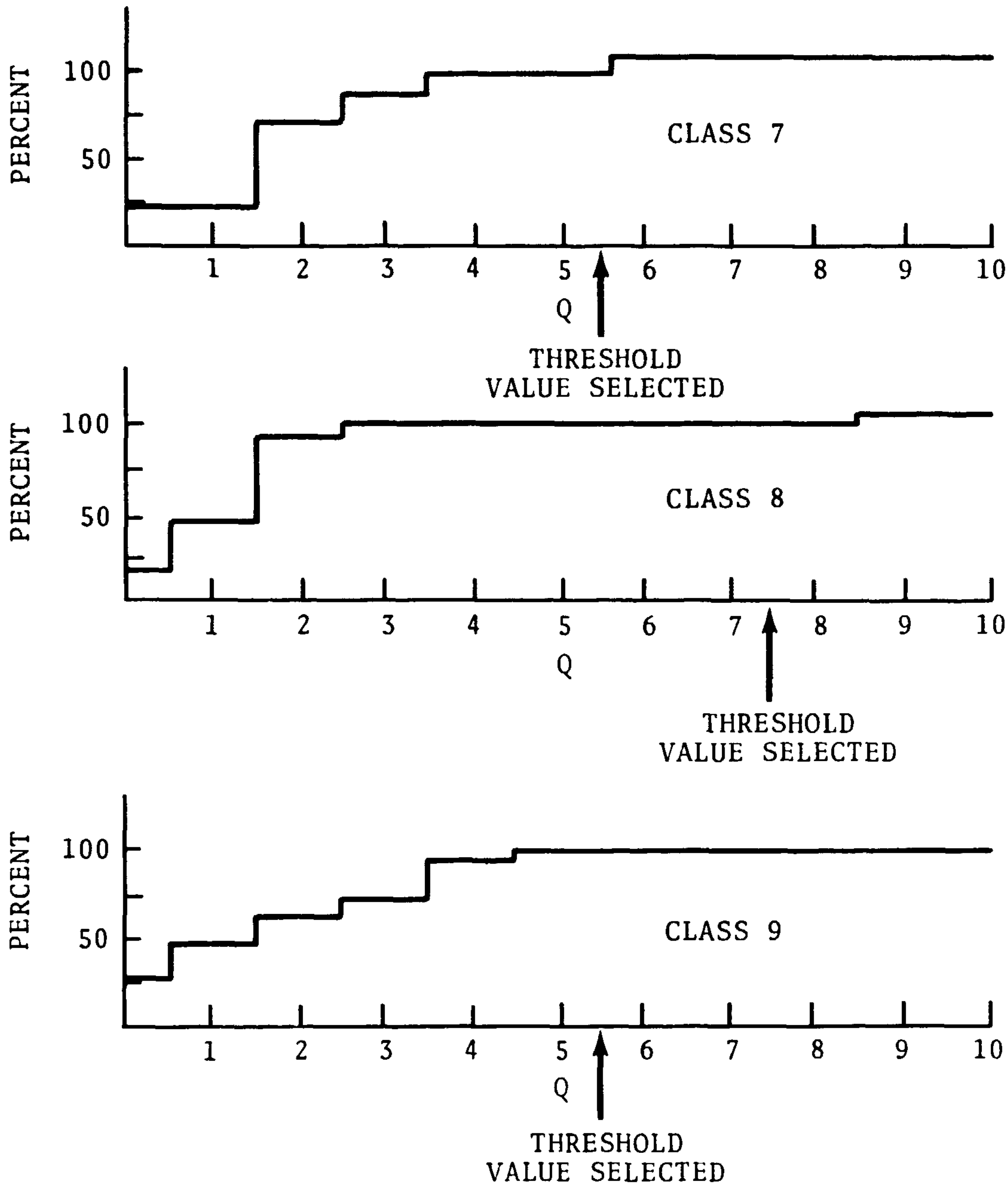


Figure 4. - Concluded.

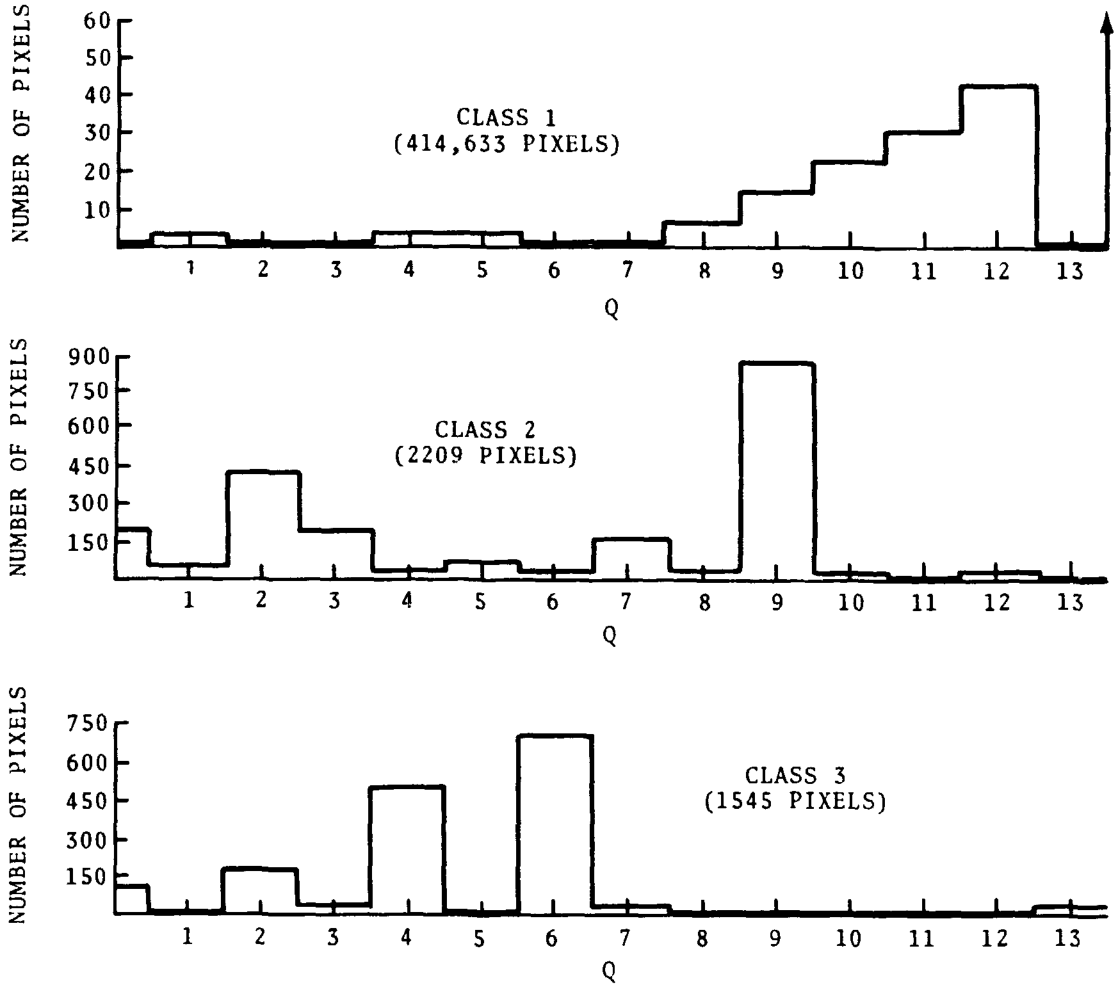


Figure 5. - Empirical estimate of the density functions of the quadratic form Q_i for the nine water classes - estimated using all of the pixels assigned to each class in the Lake Somerville study area.

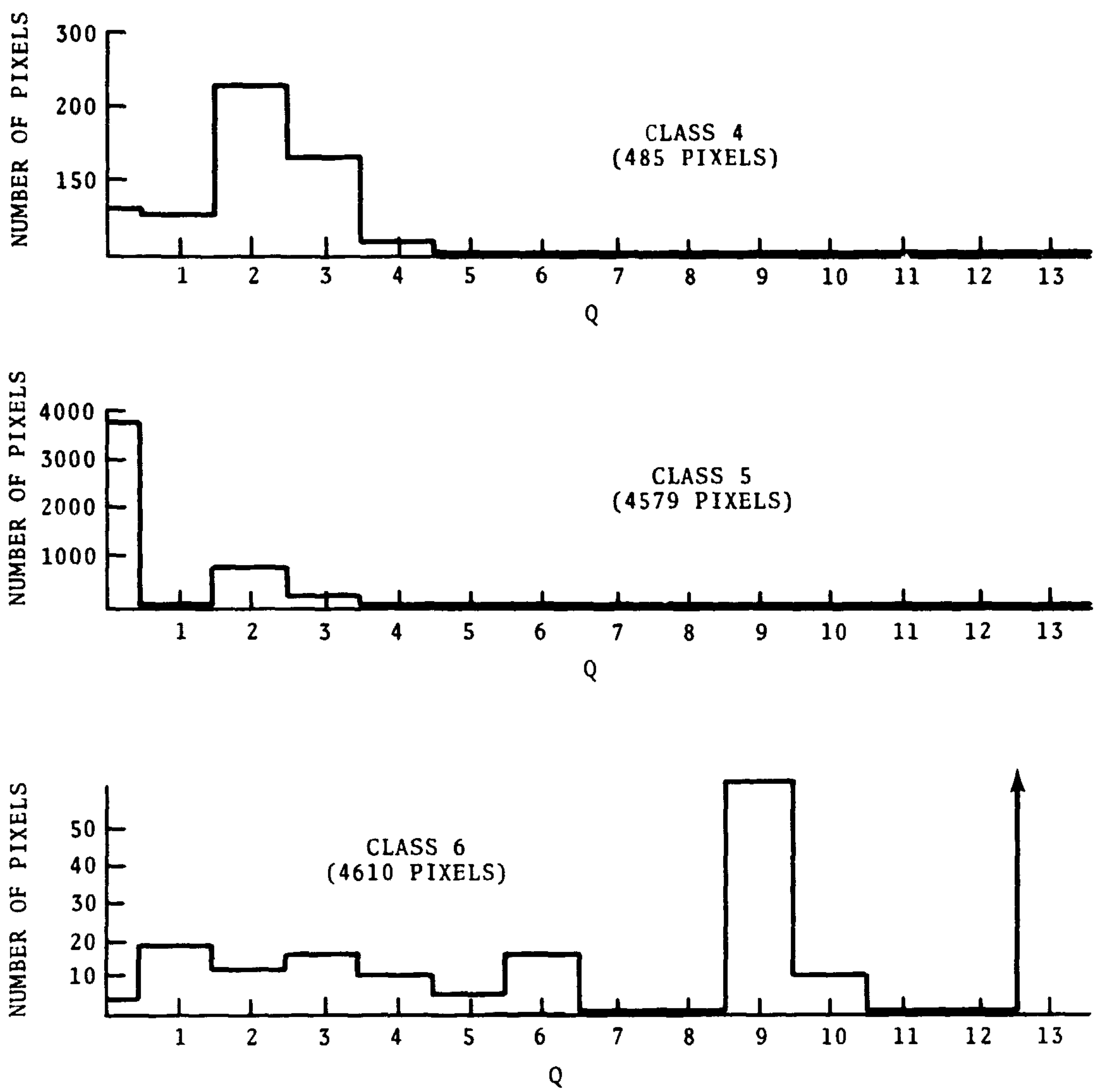


Figure 5. - Continued.

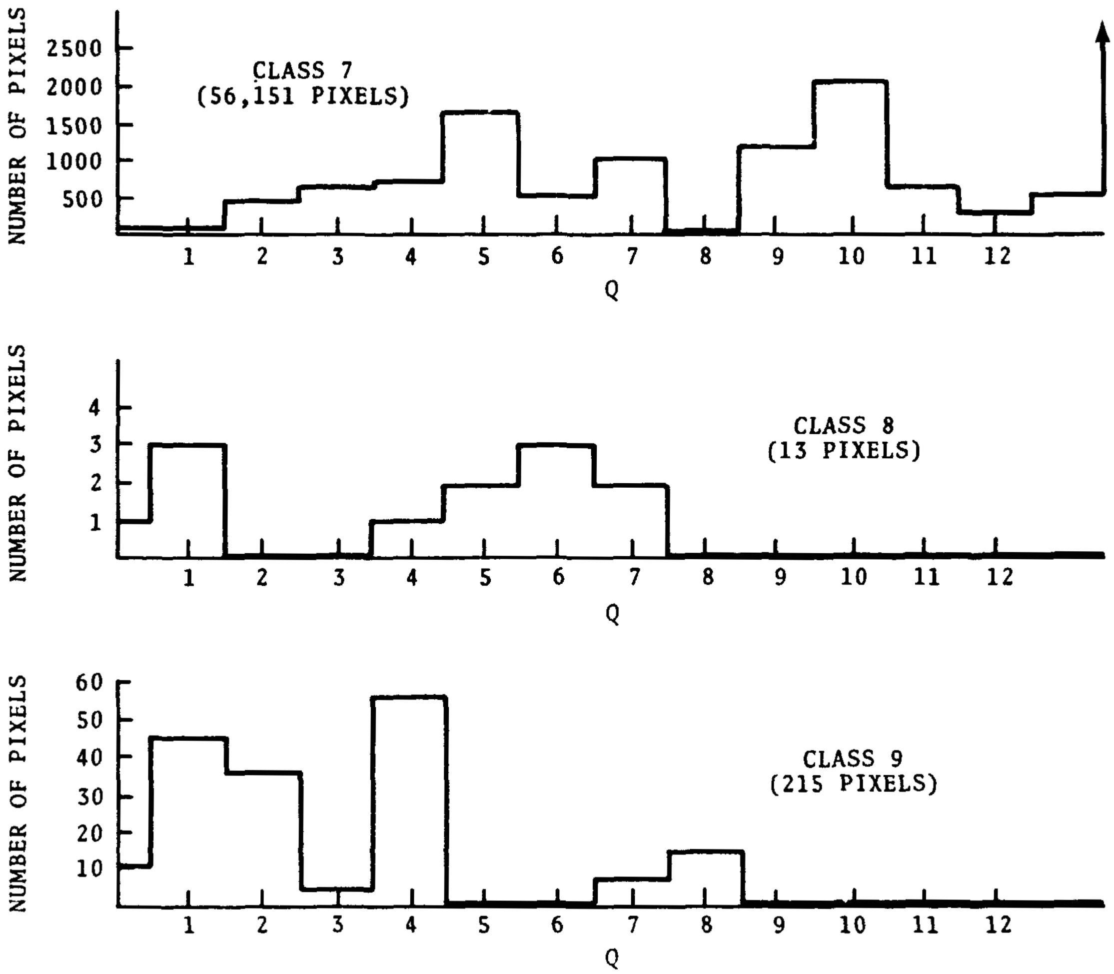
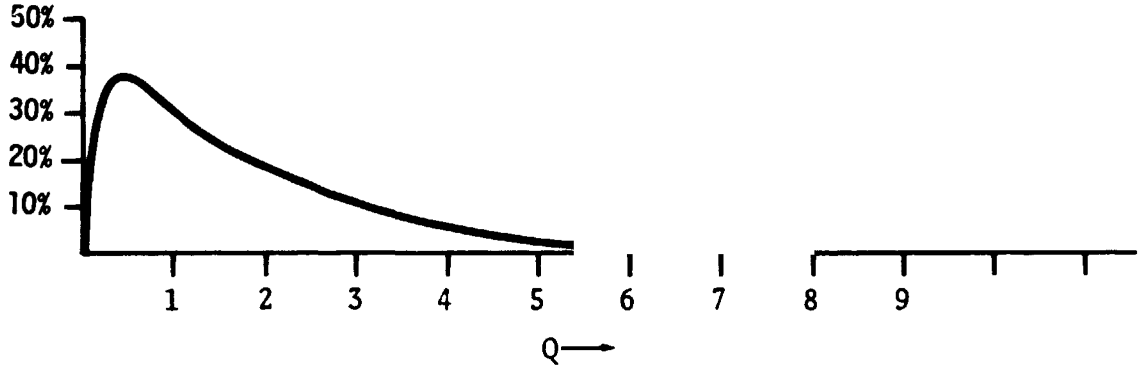
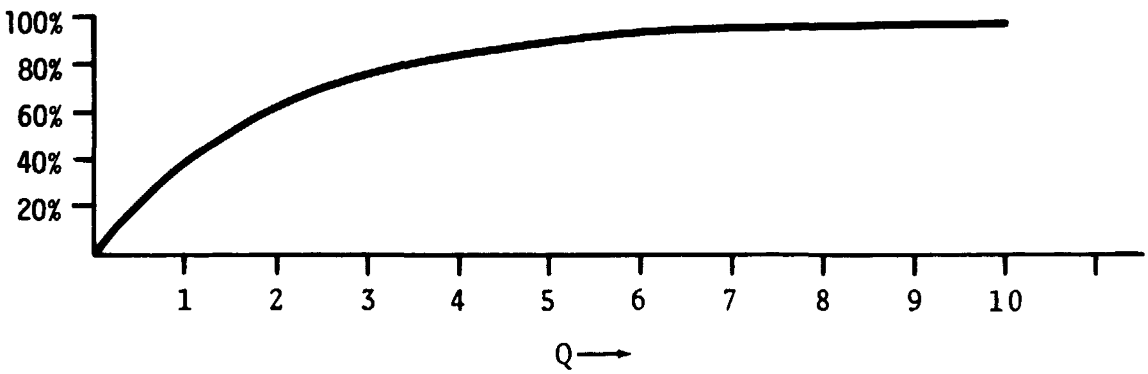


Figure 5. - Concluded.



THEORETICAL CHI-SQUARE DENSITY FUNCTION
FOR TWO CHANNELS



THEORETICAL CUMULATIVE CHI-SQUARE DISTRIBUTION
FUNCTION FOR TWO CHANNELS

Figure 6. - The theoretical chi-square density and cumulative distribution for two degrees of freedom.

approximately 3 to 5), it would not be unexpected if the tails of the density of Q_i differed significantly from the theoretical chi-square density. From the data shown in Figures 3 and 5, it can be concluded that the clusters defined by ISOCLS are generally multimodal and do not conform to the normality assumption. In addition, a comparison of the forms of the density functions of Q_i for low values of Q_i (i.e. for Q_i values less than approximately 3 to 5 in Figures 3 and 5) on a class-by-class basis indicated that the training samples for each class are not generally representative of the classes of water found in the Lake Somerville Study Area (class 5 is an example of a class for which the training samples <u>do</u> appear to be representative of that class over all the study area even though it is multimodal). Under these conditions, the procedure adopted for selecting thresholds was as follows:

(a) Using the estimate of the cumulative distribution function of Q_i obtained from the training samples and shown in Figure 4, determine what threshold value T_i (i.e. $T_i = 1/2\ Q_i$) is required to retain some initial percentage of the training samples between 90% and 100%. These bounds are imposed by the performance criteria discussed in Appendix C. An initial value of 95% was selected as the retention rate for testing this procedure.

(b) Run DISPLAY with the threshold value selected above.

(c) Evaluate the performance of the classifier using the performance evaluation criteria described in Appendix C.

(d) If the performance of the classifier is acceptable, the threshold selection procedure is completed. Otherwise go back to (a) above and pick new thresholds based on the result of the performance evaluation (i.e. if more than 90% of all areas of surface water of 10 acres or greater in size were correctly detected, with a frequency of false detection greater than 10%, then lower the training sample retention rate to a value nearer 90% or vice versa).

(F) Step 7 - DISPLAY was run with the threshold values obtained in step 6 to display the results contained on the MAPTAP obtained in step 5.

(G) Step 8 - Using the display map it was estimated that there were approximately 500 class 0 areas (i.e. non-water areas misclassified as Class III areas). The frequency of false detection was approximately 96.8%. Areas of surface water of 10 acres or more (Class III areas) were detected and located with an accuracy of 100%. Areas of surface water from 7 to 9.9 acres (Class II areas) were detected and located with an accuracy of 20%. All Class II areas were detected but many were represented by more or less than two line-printer symbols. An analysis of the display map indicated the major source of error was the misclassification of wet areas, bare soil, and perimeter pixels as water. These pixels were generally associated with classes 2 or 7. Figure 7 shows the outline of a 28.9 acre area of surface water on the display map classification results. Many of the pixels around the perimeter of the lake were identified as belonging to class 7. An analysis of five of the Class III areas indicated that when considering all of the perimeter pixels (both thresholded and unthresholded), 65% of the perimeter pixels were thresholded, 33% belonged to class 7, and 2% to class 2. Using the overlays for the perimeter of areas of surface water which were generated from photography (Mission 220), it was estimated that within the boundaries of the Class III areas (10 surface acres or more), 90% of the pixels belonged to classes 3, 4, 5, 6, and 9. Nine percent of these pixels belonged to class 7, and 1% to class 2. Of the pixels that lie on the perimeter of the areas of surface water and were not thresholded, 70% belonged to class 7, 24% to class 2, and 6% to class 1.

The mislabeling of a perimeter pixel as water does not effect the accuracy with which Class III areas are detected. However, it does increase the possibility that a Class II area will be incorrectly identified as a Class III area. Instead, the major source of error is

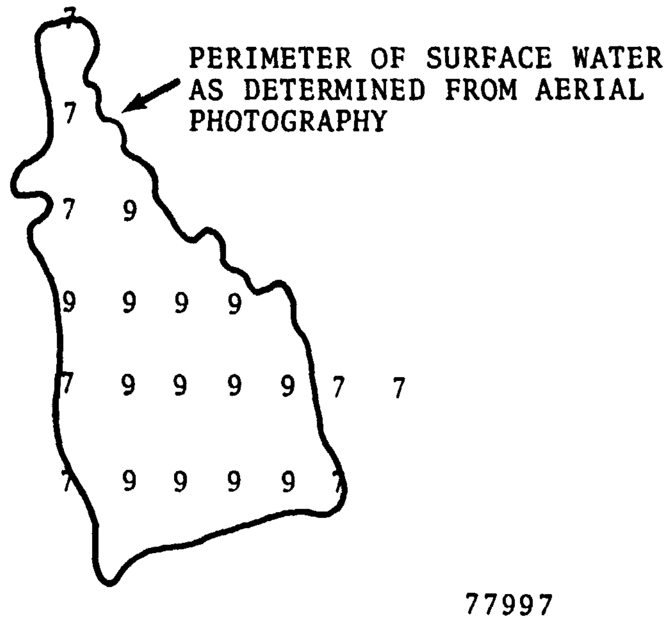


Figure 7. - Classification results for a 28.9 acre area of surface water.

the false detections which occur when wet areas and bare soil are mislabeled as water. Among the wet areas and bare soil areas identified from photography, it was found that in the wet areas 53%, of the false detections belonged to class 2, and 47% to class 7, and in the bare soil areas, 63% of the false detections belonged to class 7, and 37% to class 2. None of the other classes were present in these areas. One commercial area, which was falsely identified as water, had ? pixels from class 6, and 3 pixels from class 7.

As a result of this analysis, it was concluded that classes 2 and 7 were the major sources of false detections. To eliminate the false detections introduced by these two classes, nearly all of the pixels assigned to these two classes have to be thresholded out. If all of the pixels assigned to these two classes had been thresholded out however, one of the Class III areas would not have been correctly identified, two Class III areas would have been misclassified as Class II areas, and the remaining 13 Class III areas in the Somerville study area would have been correctly identified.

Since the preceding discussion has shown that clusters 2 and 7 were the results of incorrectly identified training fields, steps 2 and 3 of the procedure need to be changed, <u>not the thresholds</u>. Thresholding all of the pixels assigned to class 2 and 7 implies some a priori information about these two classes. This does not appear to be a reasonable approach to eliminating false detections introduced by these two classes, so no other threshold values were tried for the purpose of this study.

4.3 Conclusions

Several problem areas were noted as a result of the analysis performed during this evaluation, and warrant comment:

(1) The water training field selection procedure incorrectly included pixels from the perimeter of areas of surface water, wet fields, and bare soil areas. The presence of the wet fields/bare soil/perimeter pixels among the water training samples resulted in the presence of two clusters (clusters 2 and 7 - see Figure 2) in the non-water part of feature space. These two clusters resulted in many false detections when wet areas and bare soil were classified as water. Within the scope of this effort no practical method was developed for selecting thresholds which could be used to eliminate most of the pixels assigned to these two clusters.

(2) Another source of error can be attributed to the fact that ISOCLS, as used in this evaluation, did not define unimodal clusters and compute meaningful statistics (mainly the covariance matrix) for each class. The classes defined by ISOCLS did not appear to conform to the normality assumption and were generally multimodal. No conclusion could be reached as to the effect of this statistical problem on the performance results.

4.4 Recommendations

It is recommended that the following approach be taken in addressing the detection of surface water.

(1) Using representative water training data as derived from photographic "ground truth", define new procedures for identifying water using LARSAA: One procedure using clustering, and one procedure without. Select thresholds using the procedure described in Reference 1.

(2) Using representative water training samples and representative non-water training samples (i.e. training samples from wet areas, bare soil, vegetation, perimeter of areas of surface water, cloud and terrain shadow, etc.) that have been verified from photography, define a procedure for identifying water using LARSAA: One procedure using clustering and one procedure without. Select thresholds using the procedure described in Reference 1.

(3) Using the water and non-water training samples used to test LARSAA in item (2) above, train another available classifier (Reference 3), that is independent of some of the assumptions made in Gaussian maximum likelihood classification (such as the normality assumption and the unimodal class assumption) for the purpose of obtaining an independent assessment of how well the maximum likelihood classifier is performing in meeting its assumptions.

(4) Define a procedure for selecting representative training samples for water and non-water which do not contain mislabeled training samples.

APPENDIX A

A DETAILED DESCRIPTION OF THE EVALUATED COMPUTER-AIDED PROCESSING PROCEDURE

APPENDIX A

A DETAILED DESCRIPTION OF THE EVALUATED COMPUTER-AIDED DATA PROCESSING PROCEDURE

This procedure involved the following functional steps:

(1) Select the appropriate ERTS-1 System Corrected Computer Compatible tape.

(2) Using program REFORM, convert the tape to a format compatible with the classifier input (in this case LARSYS II).

(3) Using program PICMON, obtain a grey map of channel 4 (Caution: use the default mode for symbol selection.)

(4) Select the appropriate base maps and select ground control points. Relate these ground control points to line and column locations on the channel 4 grey map. (Reference 4)

(5) Using program REGSTR, obtain a geometrically corrected LARSYS II tape.

(6) Using program PICMON, obtain a geometrically corrected grey map of channel 4. Use the following symbols for the appropriate count range.

Symbol	Counts
m	0 thru 6
*	7 thru 9
blank	10 thru 63

(7) From the PICMON map obtained in step 6, locate all areas containing eight or more contiguously printed M or * symbols. If an area consists of at least 25% M symbols, outline this area and retain for step 8.

(8) Inscribe the largest possible rectangle, which contains at least 50% M symbols, within each of the areas located in step 7. If such a rectangle contains less than eight symbols (including M and *), delete this area from any further consideration. If the rectangle consists of more than 30 symbols, reduce the size of the rectangle so it contains no more than 30 symbols.

The procedure for selecting water training fields as outlined in steps 7 and 8 above was arbitrary.

(9) Punch LARS - 12 cards for each of the fields selected in step 8.

(10) Run ISOCLS using channels 1 and 4, with MAXCLAS = 20, NMIN = 16, STDMAX = 1.5, DLMIN = 3.2, and ISTOP = 20. Obtain a statistics deck.

(11) Run CLASSIFY, using channels 1 and 4 and the statistics for the clusters from step 10 and obtain a MAPTAP (map tape).

(12) Obtain the density and cumulative distribution of the quadratic form for each subclass of water, and select threshold values.

(13) Run DISPLY with the MAPTAP obtained in step 11 and the threshold values obtained in step 12 and obtain a line-printer output.

(14) On the display map, the water classes are displayed as integers and non-water classes are displayed as blanks.

(15) View the display map to determine the areas on the map which correspond to the following definitions:

Class I areas - A classification map area containing one ADP symbol will be defined as the ADP identification of a lake of 2 to 6.9 surface acres.

Class II areas - A classification map area containing two contiguous ADP symbols will be defined as the ADP identification of a lake of 7.0 to 9.9 surface acres.

Class III areas - A classification map area containing three or more contiguous ADP symbols will be defined as the ADP identification of a lake of 10 or more surface acres.

APPENDIX B

DETERMINING THE EMPIRICAL DISTRIBUTION OF THE QUADRATIC FORM FOR USE IN THRESHOLDING

APPENDIX B

DETERMINING THE EMPIRICAL DISTRIBUTION OF THE QUADRATIC FORM FOR USE IN THRESHOLDING

An empirical threshold selection procedure described in Reference 1 was used to select threshold values. This procedure provides a method for obtaining thresholds even when the class data does not conform to the normality assumption. This procedure will be briefly described in the following paragraphs and is essentially the same procedure as described in Reference 1.

The LARSAA classification processor, CLASSIFY, classifies measurement vectors using a maximum likelihood scheme based on an assumed multivariate normal probability density function. In the case of LARSAA, the classification and the confidence level for each pixel are stored on the output map tape by the program.

Equation 1 gives the assumed conditional probability density function for the multispectral brightness vector $X^T = (x_1, x_2, \ldots, x_i, \ldots, x_n)$ measured when material of class i fills the scanner field of view.

$$P_i(x) = \frac{1}{(2\pi)^{N/2}|K_i|^{1/2}} e^{-0.5(X - M_i)^T K_i^{-1}(X - M_i)} \tag{1}$$

Here M_i and K_i are the previously computed mean vector and covariance matrix for the i^{th} class. Because of the exponential form of equation 1 and because $\ln(P_i)$ is a monotonically increasing function of P_i it is convenient to define a new variable V_i according to equation 2.

$$V_i = \ln(P_i) \tag{2}$$

Equations 3 and 4 result from combining equations 1 and 2

$$V_i = C_i - \frac{1}{2}(X - M_i)^T K_i^{-1}(X - M_i) \tag{3}$$

where

$$C_i = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\left|K_i\right|\right) \tag{4}$$

To select the material with the maximum likelihood, the following decision rule is used:

A data vector $X^T = (x_1, x_2, \ldots, x_i, \ldots, x_N)$ is classified as belonging to class i, if,

$$V_i > V_j \text{ for all } i \neq j \tag{5}$$

Equations 3, 4, and 5 are the means by which the N measurements in X are classified as representing a particular material-type on the basis of the input mean vectors M_i and the input covariance matrices K_i.

After classification, the display processor, DISPLAY, reads and displays the results from the map tape (MAPTAP) and summarizes the results. This program has a provision for assigning samples to the unclassified category if the confidence level does not exceed a value dependent on the specified threshold T_i; this condition is given by equation 6. If

$$P_i(X) < \left|\max(P_i)\right| \quad e^{-T_i} \tag{6}$$

then X is not assigned to a category, where $\max(P_i)$, given by equation 7, is the maximum value of P_i.

$$\max(P_i) = \max_i P_i(M_i) = \max_i \exp(V_i) \tag{7}$$

By combining equations 1, 6, and 7 the condition for assigning a sample to the unclassified category can be expressed by equation 8.

$$Q_i(x) = (X - M_i)^T K_i^{-1}(X - M_i) > 2T_i \tag{8}$$

The value of the quadratic form, Q_i, is often called Mahalanobis' Distance; it is the weighted (by K_i^{-1}) squared distance in N-space between a given measurement vector and the mean vector for class i. According to equation 8, a sample is left unclassified if its Mahalanobis' Distance exceeds the value $2T_i$.

In using the LARSAA program, DISPLAY, the investigator can specify the threshold value, T_i, for all materials. In the past this was done in two ways.

(1) The DISPLAY program was used with various values of T_i to obtain a classification map (in the form of a line-printer listing) which was subjectively determined to be adequate. This meant that most of the training and test fields were correctly classified and yet almost no classifications were made where the material type was known to differ from those materials defined by input mean vectors and covariance matrixes.

(2) It is well known that in theory if P(x) is normal then the quantity Q_i in equation 8 is a random variable having a chi-square distribution with N degrees of freedom (where N is the dimension of the measurement vector X). To the extent that the training data is normally distributed, the investigator can look up in a cumulative chi-square table the threshold value which will reject (i.e., assign to the unclassified category) no more than a specified percentage of samples. For example, a threshold setting of T_i = 3.0 will reject no more than 5 percent of samples drawn from a two-dimensional multivariate normal distribution.

Since the training data is not normally distributed in many cases, the distribution of Q_i must be determined empirically to select the threshold value. To empirically determine the thresholds for the nine water classes, the LARSAA map tape (MAPTAP) generated in classification together with a card deck giving ground truth information for the retangular water training fields were input to the Density and Distribution Function Program to compute the actual distribution of Q_i. Appendix D contains a listing of the Density and Distribution Program. The LARSAA (MAPTAP) contained the class constants C_i for each class, and for each pixel:

(1) The integer value for i, the most likely material.

(2) The floating-point variable V_i define by equations 2 and 1 for each of the 439 elements in each of the 1100 lines on the map tape.

Figure B1 shows the flow of information involved in computing the density and distribution functions. For each scan line, 439 values of i and V_i are read into core. Each element is considered to determine whether or not ground truth is available (i.e., if it lies within one of the rectangular fields defined by the Ground Truth deck). If a pixel lay within one of the rectangular training fields defined by the Ground Truth deck, the value V_i is used with the class constant (C_i for that particular material) to determine Q_i the Mahalanobis' Distance. The number of occurrences for each value of Q_i for each material are accumulated to form an estimate of the <u>density</u> function. After all samples on the input map tape have been tested, the density functions are accumulated (i.e., summed) with respect to Q_i to form the <u>distribution</u> function for each material. The empirical density and distribution functions are both printed out as functions of Q_i for each of the nine materials.

Figure B2 shows the flow of information involved in computing the density and distribution functions when the map tape contains only the classified water training fields. This modification to the Density and Distribution Functions Program allowed water training fields which were outside the Lake Somerville Study Area to be included in the computation of the density and distribution for the quadratic form $Q_i(X)$ for each water class.

In addition, the Figure B2 flow was used to obtain the density function and distribution function for the quadratic form Q_i for each class using all of the pixels in the Somerville study area which were assigned to each class. To accomplish this, the whole Lake Somerville Study Area was treated as a single water training field in the Figure B2 flow.

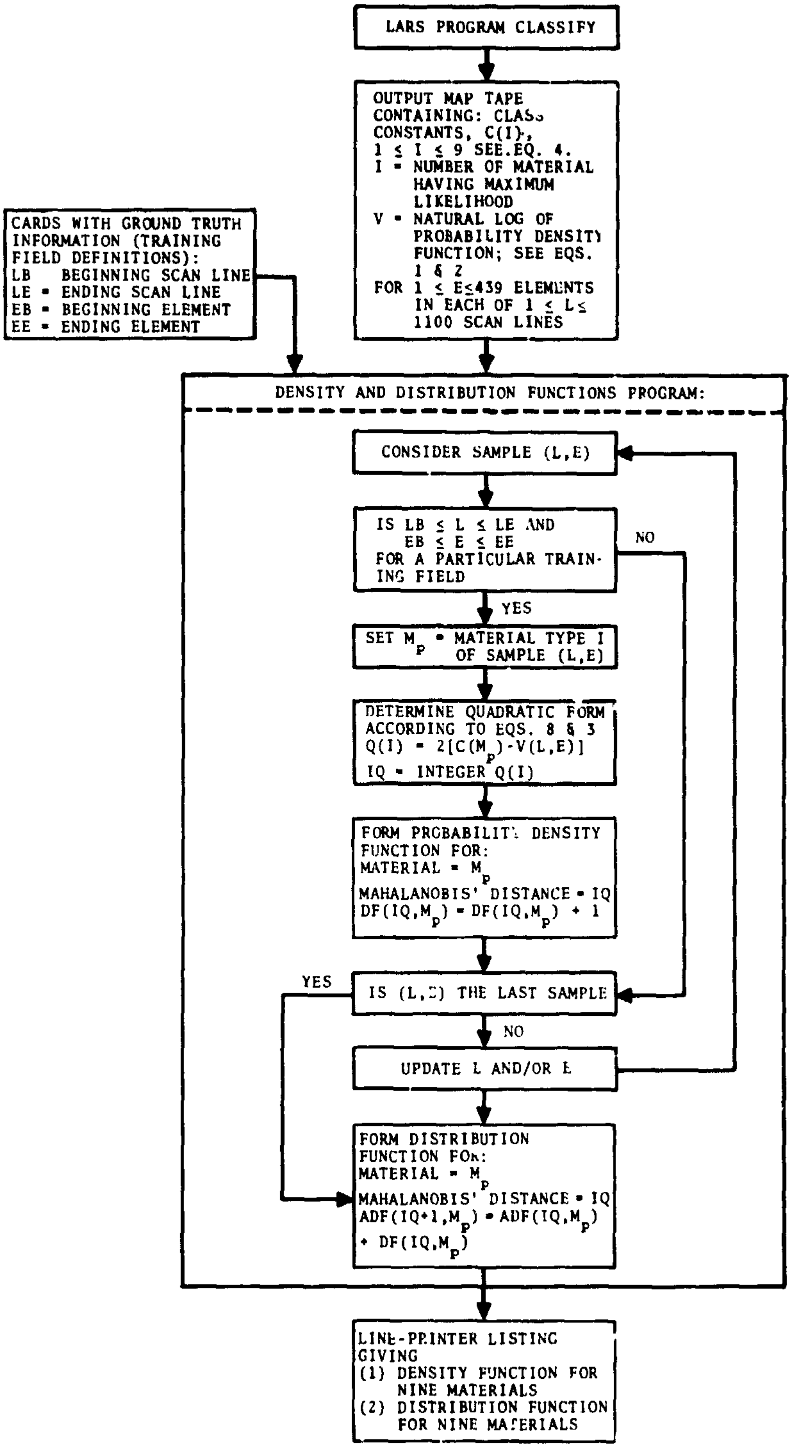


Figure B1. - Flow diagram for computing density and distribution functions for the general case when training fields lie within the bounds of the area classified on the MAPTAP.

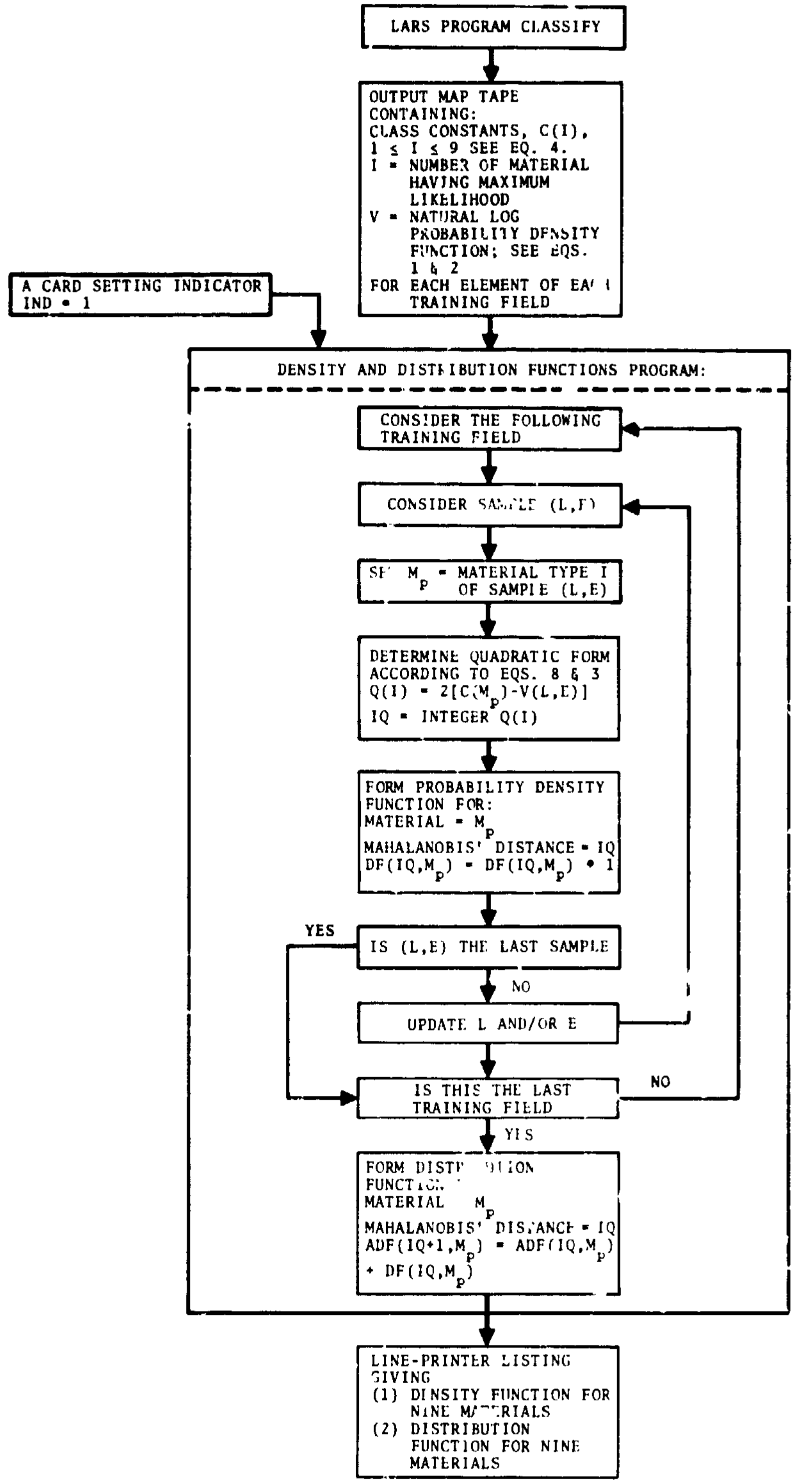


Figure B2. - Flow diagram for computing density and distribution functions when MAPTAP contains only the classified water training fields.

APPENDIX C

A DETAILED DESCRIPTION OF THE PERFORMANCE EVALUATION PROCEDURE

APPENDIX C

A DETAILED DESCRIPTION OF THE PERFORMANCE EVALUATION PROCEDURE

The performance of the procedure described in Section 4.1 was evaluated using photointerpretation data along with the following evaluation rationale. The procedure is acceptable if it meets or exceeds the following criteria.

(a) Detect and locate all Class III areas with an accuracy of 90% or greater.

(b) Frequency of false detections of 10% or less on Class III areas.

Items (a) and (b) correspond to F_{33} and F_{03} of the matrix below. The remaining members of this matrix were not evaluated for this procedure.*

*This has been done for other procedures tested. (See Reference 2.)

PERFORMANCE MATRIX

True Class	Assigned Class I	Assigned Class II	Assigned Class III
I	F_{11}	F_{12}	F_{13}
II	F_{21}	F_{22}	F_{23}
III	F_{31}	F_{32}	F_{33}
0	F_{01}	F_{02}	F_{03}

F_{ij} = frequency with which ADP identification of class j areas were actually class i areas (i.e. class j areas mis-identified as class i areas).

F_{oj} = frequency with which ADP identifications of class j areas were actually not an area of any class. (Frequency of False Detection)

$F_{ii} = \frac{L_i}{M_i}$ $\qquad$ $F_{ij} = \frac{K_{ij}}{N_j}$

L_i = the total number of correct ADP identifications of class i areas.

M_i = the total number of class i areas in study area.

N_j = the total number class j areas.

K_{ij} = number of ADP identifications of class j areas which were actually class i areas.

K_{oj} = number of ADP identifications of class j areas which were actually not an area of any class.

APPENDIX D

THE DENSITY AND DISTRIBUTION COMPUTER PROGRAM LISTING

APPENDIX D

THE DENSITY AND DISTRIBUTION PROGRAM COMPUTER LISTING

The following program was used to compute the density and distribution of the quadratic form Q(x). It is written in FORTRAN IV and uses approximately 500 words of core storage for code and approximately 6,000 words for data. To process 20 training fields of water it takes approximately 30 seconds of CPU time on the Univac 1108 EXEC 2.

```
      PROGRAM COMPUTES DISTRIBUTION VS DISTANCE FOR THRESHOLD SETTING
      DIMENSION CONS(24),IA(1000),V(1000),NO(24,100), NP(100),DF(100),
     1DFA(100),         LB(50),LE(50),NSB(50),NSE(50),NTRY(24),IB(1000)
      INTEGER VARSZ2
      DIMENSION COVMTX(300,30)
  100 FORMAT(4I10)
  101 FORMAT(20X,2I5,5X,2I5)
  102 FORMAT(1H1,10I13)
  103 FORMAT(8F10.2)
  104 FORMAT(1H0,8F10.2)
  105 FORMAT(1H0,10I13)
  106 FORMAT(1H0,10E13.4)
  107 FORMAT(1H0,4I10)
  108 FORMAT(1H1,3I10,1F10.1)
      READ(1) SERIAL,NCLS,NOFLD2,VARSZ2
      READ(1) DUM
      READ(1) DUM
      READ(1) ((COVMTX(I,J),I=1,VARSZ2),J=1,NCLS),(CONS(I),I=1,NCLS)
      READ(5,100) NLB,NLE,NTST,IND
      WRITE(6,102)NCLS,NLB,NLE,NTST
      WRITE(6,104)(CONS(I),I=1,NCLS)
      IF(IND.NE.0) GO TO 50
      READ(5,101) (           LB(I),LE(I),NSB(I),NSE(I)  ,I=1,NTST)
      WRITE(6,107)(           LB(I),LE(I),NSB(I),NSE(I)  ,I=1,NTST)
   50 CONTINUE
      DO 22 I=1,NCLS
      NTRY(I)=0
   22 CONTINUE
      M=0
      NT=0
      DO 1 I=1,NCLS
      DO 2 J=1,100
      NO(I,J)=0
    2 CONTINUE
    1 CONTINUE
    3 CONTINUE
      READ(1) IPTS,LINES
      IF(IPTS.EQ.0) GO TO 6
      READ(1) L,(IA(K),K=1,IPTS),(V(K),K=1,IPTS)
      IF(IND.NE.0) GO TO 51
      IF(L.LT.NLB) GO TO 3
    4 CONTINUE
      DO 30 K=1,IPTS
      IB(K)=0
   30 CONTINUE
      DO 31 N=1,NTST
      IF(L.LT.LB(N).OR.L.GT.LE(N)) GO TO 31
```

```
C*    MATP=MAT(N)
      ISB=NSB(N)
      ISE=NSE(N)
      DO 5 K=ISB,ISE
      IB(K)=IA(K)
    5 CONTINUE
   31 CONTINUE
   51 CONTINUE
      DO 32 K=1,IPTS
      M=IA(K)
      IF(IND.NE.0) GO TO 52
      N=IB(K)
      IF(M.NE.N) GO TO 33
   52 CONTINUE
      Q=2.0*(CONS(M)-V(K))+1.5
      IQ=INT(Q)
      IF(IQ.GT.0) GO TO 11
      IQ=99
      GO TO 12
   11 CONTINUE
      IF(IQ.LT.99) GO TO 12
      IQ=100
   12 CONTINUE
      NO(M,IQ)=NO(M,IQ)+1
   33 CONTINUE
   32 CONTINUE
      IF(IND.NE.0) GO TO 53
      IF(L.EQ.NLE) GO TO 6
   53 CONTINUE
      READ(1) L,(IA(K),K=1,IPTS),(V(K),K=1,IPTS)
      IF(IND.NE.0.AND.L.EQ.0) GO TO 3
      IF(L.EQ.0) GO TO 6
      IF(IND.NE.0) GO TO 51
      GO TO 4
    6 CONTINUE
      DO 7 M=1,NCLS
      NT=0
      DO 8 J=1,100
      N=NO(M,J)
      NP(J)=N
      NT=NT+N
    8 CONTINUE
      B=FLOAT(NT)
      DA=0.0
      DO 9 J=1,100
      N=NP(J)
```

```
      T=FLOAT(N)
      D=T/B
      DF(J)=D
      DA=DA+D
      DFA(J)=DA
    9 CONTINUE
      WRITE(6,108) M,NT
      WRITE(6,105) NP
      WRITE(6,108) M,NT
      WRITE(6,105) DF
      WRITE(6,108) M,NT
      WRITE(6,105) DFA
    7 CONTINUE
      STOP
      END
```

REFERENCES

1. Eppler, W. B., "Emperical Distribution of Quadratic Form Used for Thresholding", LEC/HASD #FSD-0001, Lockheed Electronics Company, Houston Aerospace Systems Division, Houston, Texas, November 1972.

2. Moore, B. H., "Development of a Computer-Aided Procedure for the National Program for Inspection of Dams", JSC-08449, Johnson Space Center, Houston, Texas, August 1973.

3. Pitt, J. M., "An Adaptive Multicategory Pattern Classification System", Ph.D. Dissertation, Laboratories for Electronics and Related Science Research, Technical Report #43, The University of Texas at Austin, January 22, 1968.

4. "Procedures Manual for Detection and Location of Surface Water Using ERTS-1 Multispectral Scanner Data: Vol. III, Control Network Establishment," JSC-08451, Johnson Space Center, Houston, Texas, November 1973.